EXHIBIT HH

How Tech Giants Cut Corners to Harvest Data for A.I.

OpenAI, Google and Meta ignored corporate policies, altered their own rules and discussed skirting copyright law as they sought online information to train their newest artificial intelligence systems.

By Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson and Nico Grant

Reporting from San Francisco, Washington and New York

Published April 6, 2024 Updated April 8, 2024

In late 2021, OpenAI faced a supply problem.

The artificial intelligence lab had exhausted every reservoir of reputable English-language text on the internet as it developed its latest A.I. system. It needed more data to train the next version of its technology — lots more.

So OpenAI researchers created a speech recognition tool called Whisper. It could transcribe the audio from YouTube videos, yielding new conversational text that would make an A.I. system smarter.

Some OpenAI employees discussed how such a move might go against YouTube's rules, three people with knowledge of the conversations said. YouTube, which is owned by Google, prohibits use of its videos for applications that are "independent" of the video platform.

Ultimately, an OpenAI team transcribed more than one million hours of YouTube videos, the people said. The team included Greg Brockman, OpenAI's president, who personally helped collect the videos, two of the people said. The texts were then fed into a system called GPT-4, which was widely considered one of the world's most powerful A.I. models and was the basis of the latest version of the ChatGPT chatbot.

The race to lead A.I. has become a desperate hunt for the digital data needed to advance the technology. To obtain that data, tech companies including OpenAI, Google and Meta have cut corners, ignored corporate policies and debated bending the law, according to an examination by The New York Times.

At Meta, which owns Facebook and Instagram, managers, lawyers and engineers last year discussed buying the publishing house Simon & Schuster to procure long works, according to recordings of internal meetings obtained by The Times. They also conferred on gathering copyrighted data from across the internet, even if that meant facing lawsuits. Negotiating licenses with publishers, artists, musicians and the news industry would take too long, they said.

Like OpenAI, Google transcribed YouTube videos to harvest text for its A.I. models, five people with knowledge of the company's practices said. That potentially violated the copyrights to the videos, which belong to their creators.

Last year, Google also broadened its terms of service. One motivation for the change, according to members of the company's privacy team and an internal message viewed by The Times, was to allow Google to be able to tap publicly available Google Docs, restaurant reviews on Google Maps and other online material for more of its A.I. products.

The companies' actions illustrate how online information — news stories, fictional works, message board posts, Wikipedia articles, computer programs, photos, podcasts and movie clips — has increasingly become the lifeblood of the booming A.I. industry. Creating innovative systems depends on having enough data to teach the technologies to instantly produce text, images, sounds and videos that resemble what a human creates.

The volume of data is crucial. Leading chatbot systems have learned from pools of digital text spanning as many as three trillion words, or roughly twice the number of words stored in Oxford University's Bodleian Library, which has collected manuscripts since 1602. The most prized data, A.I. researchers said, is high-quality information, such as published books and articles, which have been carefully written and edited by professionals.

For years, the internet — with sites like Wikipedia and Reddit — was a seemingly endless source of data. But as A.I. advanced, tech companies sought more repositories. Google and Meta, which have billions of users who produce search queries and social media posts every day, were largely limited by privacy laws and their own policies from drawing on much of that content for A.I.

Their situation is urgent. Tech companies could run through the high-quality data on the internet as soon as 2026, according to Epoch, a research institute. The companies are using the data faster than it is being produced.

“The only practical way for these tools to exist is if they can be trained on massive amounts of data without having to license that data,” Sy Damle, a lawyer who represents Andreessen Horowitz, a Silicon Valley venture capital firm, said of A.I. models last year in a public discussion about copyright law. “The data needed is so massive that even collective licensing really can’t work.”

Tech companies are so hungry for new data that some are developing “synthetic” information. This is not organic data created by humans, but text, images and code that A.I. models produce — in other words, the systems learn from what they themselves generate.

OpenAI said each of its A.I. models “has a unique data set that we curate to help their understanding of the world and remain globally competitive in research.” Google said that its A.I. models “are trained on some YouTube content,” which was allowed under agreements with YouTube creators, and that the company did not use data from office apps outside of an experimental program. Meta said it had “made aggressive investments” to integrate A.I. into its services and had billions of publicly shared images and videos from Instagram and Facebook for training its models.

For creators, the growing use of their works by A.I. companies has prompted lawsuits over copyright and licensing. The Times sued OpenAI and Microsoft last year for using copyrighted news articles without permission to train A.I. chatbots. OpenAI and Microsoft have said using the articles was “fair use,” or allowed under copyright law, because they transformed the works for a different purpose.

More than 10,000 trade groups, authors, companies and others submitted comments last year about the use of creative works by A.I. models to the Copyright Office, a federal agency that is preparing guidance on how copyright law applies in the A.I. era.

Justine Bateman, a filmmaker, former actress and author of two books, told the Copyright Office that A.I. models were taking content — including her writing and films — without permission or payment.

“This is the largest theft in the United States, period,” she said in an interview.

‘Scale Is All You Need’

Jared Kaplan, a theoretical physicist at Johns Hopkins University, wrote a key paper on A.I. and data. He is also the chief science officer of the A.I. start-up Anthropic. Chris J. Ratcliffe/Bloomberg

In January 2020, Jared Kaplan, a theoretical physicist at Johns Hopkins University, published a groundbreaking paper on A.I. that stoked the appetite for online data.

His conclusion was unequivocal: The more data there was to train a large language model — the technology that drives online chatbots — the better it would perform. Just as a student learns more by reading more books, large language models can better pinpoint patterns in text and be more accurate with more information.

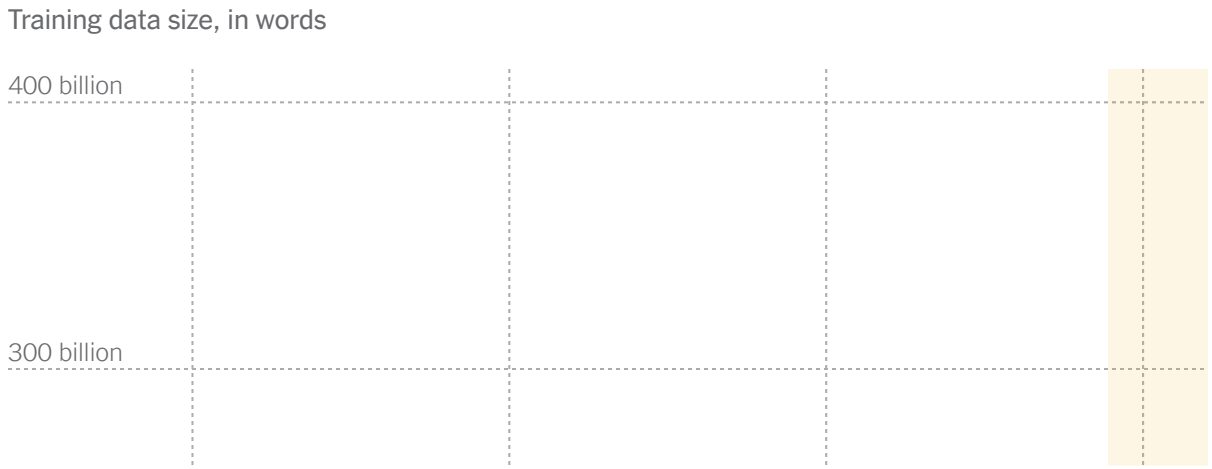
“Everyone was very surprised that these trends — these scaling laws as we call them — were basically as precise as what you see in astronomy or physics,” said Dr. Kaplan, who published the paper with nine OpenAI researchers. (He now works at the A.I. start-up Anthropic.)

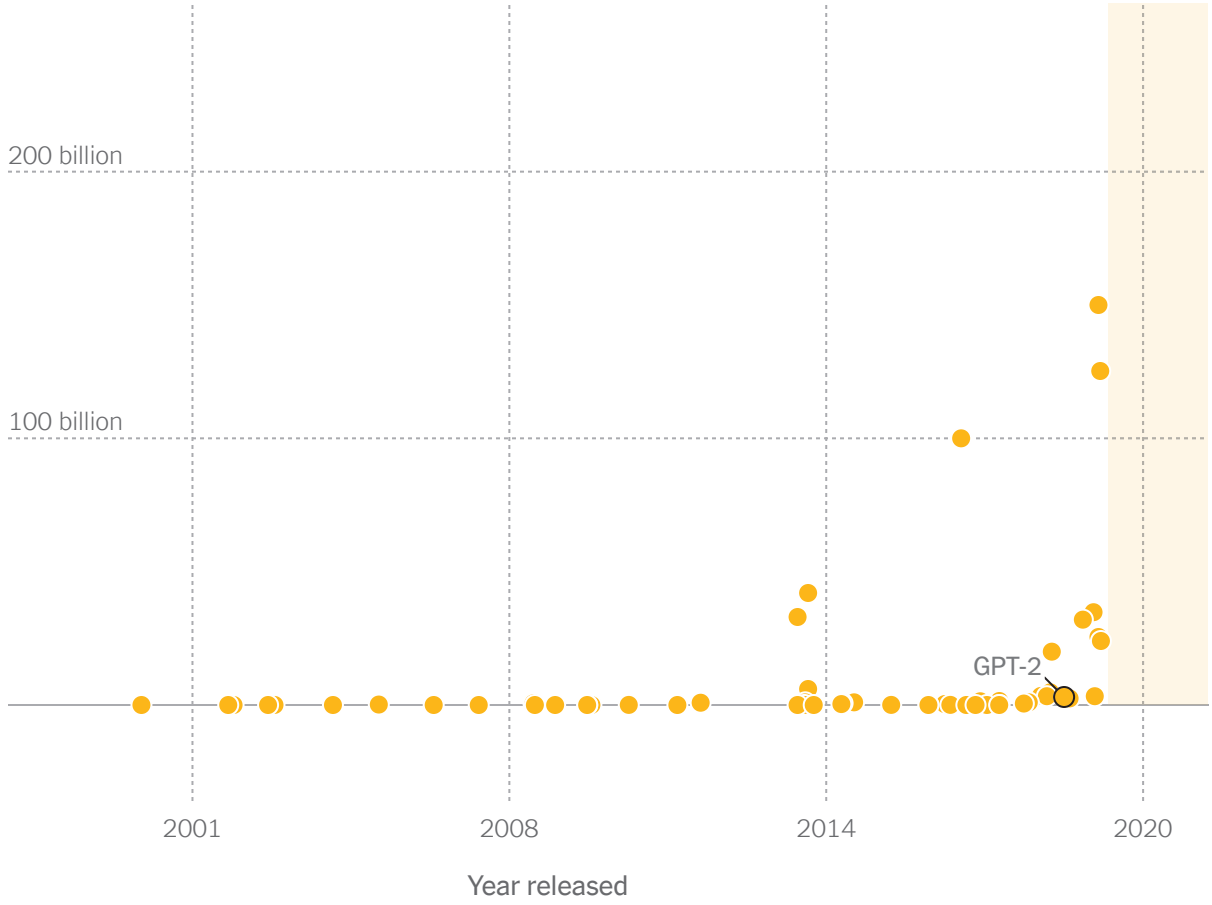
“Scale is all you need” soon became a rallying cry for A.I.

Researchers have long used large public databases of digital information to develop A.I., including Wikipedia and Common Crawl, a database of more than 250 billion web pages collected since 2007. Researchers often “cleaned” the data by removing hate speech and other unwanted text before using it to train A.I. models.

In 2020, data sets were tiny by today’s standards. One database containing 30,000 photographs from the photo website Flickr was considered a vital resource at the time.

After Dr. Kaplan’s paper, that amount of data was no longer enough. It became all about “just making things really big,” said Brandon Duderstadt, the chief executive of Nomic, an A.I. company in New York.




200 billion
100 billion
GPT-2
2001
2008
2014
2020
Year released

Before 2020, most **A.I. models** used relatively little training data.

Mr. Kaplan's paper, released in 2020, led to a **new era** defined by GPT-3, a large language model, where researchers began including more data in their models …

… much, much more data.

Note: Includes estimates. Source: Epoch.

When OpenAI unveiled GPT-3 in November 2020, it was trained on the largest amount of data to date — about 300 billion "tokens," which are essentially words or pieces of words. After learning from that data, the system generated text with astounding accuracy, writing blog posts, poetry and its own computer programs.

In 2022, DeepMind, an A.I. lab owned by Google, went further. It tested 400 A.I. models and varied the amount of training data and other factors. The top-performing models used even more data than Dr. Kaplan had predicted in his paper. One model, Chinchilla, was trained on 1.4 trillion tokens.

It was soon overtaken. Last year, researchers from China released an A.I. model, Skywork, which was trained on 3.2 trillion tokens from English and Chinese texts. Google also unveiled an A.I. system, PaLM 2, which topped 3.6 trillion tokens.

Transcribing YouTube

In May, Sam Altman, the chief executive of OpenAI, acknowledged that A.I. companies would use up all viable data on the internet.

“That will run out,” he said in a speech at a tech conference.

Mr. Altman had seen the phenomenon up close. At OpenAI, researchers had gathered data for years, cleaned it and fed it into a vast pool of text to train the company’s language models. They had mined the computer code repository GitHub, vacuumed up databases of chess moves and drawn on data describing high school tests and homework assignments from the website Quizlet.

By late 2021, those supplies were depleted, said eight people with knowledge of the company, who were not authorized to speak publicly.

OpenAI was desperate for more data to develop its next-generation A.I. model, GPT-4. So employees discussed transcribing podcasts, audiobooks and YouTube videos, the people said. They talked about creating data from scratch with A.I. systems. They also considered buying start-ups that had collected large amounts of digital data.

OpenAI eventually made Whisper, the speech recognition tool, to transcribe YouTube videos and podcasts, six people said. But YouTube prohibits people from not only using its videos for “independent” applications, but also accessing its videos by “any automated means (such as robots, botnets or scrapers).”

OpenAI employees knew they were wading into a legal gray area, the people said, but believed that training A.I. with the videos was fair use. Mr. Brockman, OpenAI’s president, was listed in a research paper as a creator of Whisper. He

personally helped gather YouTube videos and fed them into the technology, two people said.

Mr. Brockman referred requests for comment to OpenAI, which said it uses "numerous sources" of data.

Last year, OpenAI released GPT-4, which drew on the more than one million hours of YouTube videos that Whisper had transcribed. Mr. Brockman led the team that developed GPT-4.

Some Google employees were aware that OpenAI had harvested YouTube videos for data, two people with knowledge of the companies said. But they didn't stop OpenAI because Google had also used transcripts of YouTube videos to train its A.I. models, the people said. That practice may have violated the copyrights of YouTube creators. So if Google made a fuss about OpenAI, there might be a public outcry against its own methods, the people said.

Matt Bryant, a Google spokesman, said the company had no knowledge of OpenAI's practices and prohibited "unauthorized scraping or downloading of YouTube content." Google takes action when it has a clear legal or technical basis to do so, he said.

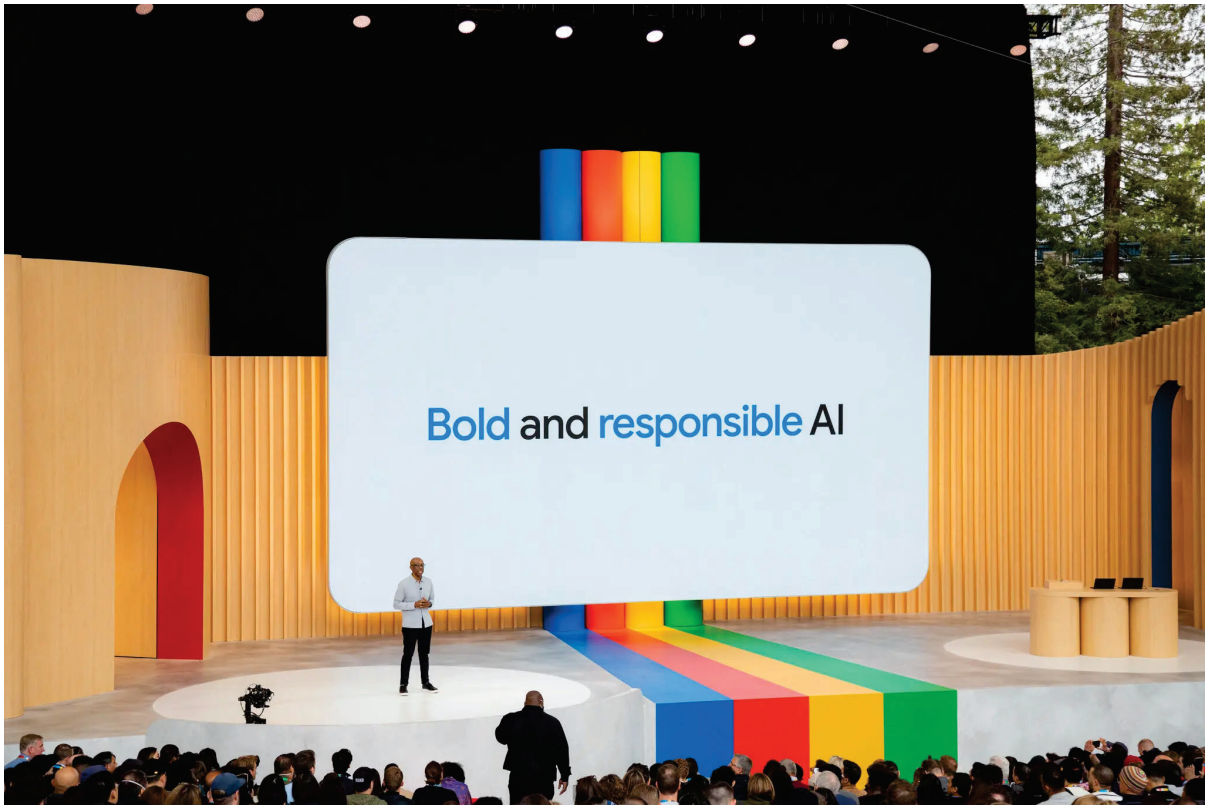
Google's rules allowed it to tap YouTube user data to develop new features for the video platform. But it was unclear whether Google could use YouTube data to build a commercial service beyond the video platform, such as a chatbot.

Geoffrey Lottenberg, an intellectual property lawyer with the law firm Berger Singerman, said Google's language about what it could and could not do with YouTube video transcripts was vague.

"Whether the data could be used for a new commercial service is open to interpretation and could be litigated," he said.

In late 2022, after OpenAI released ChatGPT and set off an industrywide race to catch up, Google researchers and engineers discussed tapping other user data. Billions of words sat in people's Google Docs and other free Google apps. But the

company's privacy restrictions limited how they could use the data, three people with knowledge of Google's practices said.



After OpenAI released ChatGPT, Google researchers and engineers discussed tapping other user data to develop A.I. products, people with knowledge of the discussions said. Jason Henry for The New York Times

In June, Google's legal department asked the privacy team to draft language to broaden what the company could use consumer data for, according to two members of the privacy team and an internal message viewed by The Times.

The employees were told Google wanted to use people's publicly available content in Google Docs, Google Sheets and related apps for an array of A.I. products. The employees said they didn't know if the company had previously trained A.I. on such data.

At the time, Google's privacy policy said the company could use publicly available information only to "help train Google's language models and build features like Google Translate."

The privacy team wrote new terms so Google could tap the data for its "A.I. models and build products and features like Google Translate, Bard and Cloud AI capabilities," which was a wider collection of A.I. technologies.

"What is the end goal here?" one member of the privacy team asked in an internal message. "How broad are we going?"

The team was told specifically to release the new terms on the Fourth of July weekend, when people were typically focused on the holiday, the employees said. The revised policy debuted on July 1, at the start of the long weekend.

How Google Can Use Your Data

Here are the changes Google made to its privacy policy last year for its free consumer apps.

Google uses information to improve our services and to develop new products, features and technologies that benefit our users and the public. For example, we use publicly available information to help train Google's ~~*language*~~ **AI** models and build **products and** features like Google Translate **, Bard, and Cloud AI capabilities** .

Source: Google • By The New York Times

In August, two privacy team members said, they pressed managers on whether Google could start using data from free consumer versions of Google Docs, Google Sheets and Google Slides. They were not given clear answers, they said.

Mr. Bryant said that the privacy policy changes had been made for clarity and that Google did not use information from Google Docs or related apps to train language models "without explicit permission" from users, referring to a voluntary program that allows users to test experimental features.

"We did not start training on additional types of data based on this language change," he said.

The Debate at Meta

Mark Zuckerberg, Meta's chief executive, had invested in A.I. for years — but suddenly found himself behind when OpenAI released ChatGPT in 2022. He immediately pushed to match and exceed ChatGPT, calling executives and engineers at all hours of the night to push them to develop a rival chatbot, said three current and former employees, who were not authorized to discuss confidential conversations.

But by early last year, Meta had hit the same hurdle as its rivals: not enough data.

Ahmad Al-Dahle, Meta's vice president of generative A.I., told executives that his team had used almost every available English-language book, essay, poem and news article on the internet to develop a model, according to recordings of internal meetings, which were shared by an employee.

Meta could not match ChatGPT unless it got more data, Mr. Al-Dahle told colleagues. In March and April 2023, some of the company's business development leaders, engineers and lawyers met nearly daily to tackle the problem.

Some debated paying $10 a book for the full licensing rights to new titles. They discussed buying Simon & Schuster, which publishes authors like Stephen King, according to the recordings.

They also talked about how they had summarized books, essays and other works from the internet without permission and discussed sucking up more, even if that meant facing lawsuits. One lawyer warned of "ethical" concerns around taking intellectual property from artists but was met with silence, according to the recordings.

Mr. Zuckerberg demanded a solution, employees said.

"The capability that Mark is looking for in the product is just something that we currently aren't able to deliver," one engineer said.

Mark Zuckerberg, Meta's chief executive, pushed his company to catch up in generative A.I. after OpenAI released ChatGPT. Jason Andrew for The New York Times

While Meta operates giant social networks, it didn't have troves of user posts at its disposal, two employees said. Many Facebook users had deleted their earlier posts, and the platform wasn't where people wrote essay-type content, they said.

Meta was also limited by privacy changes it introduced after a 2018 scandal over sharing its users' data with Cambridge Analytica, a voter-profiling company.

Mr. Zuckerberg said in a recent investor call that the billions of publicly shared videos and photos on Facebook and Instagram are "greater than the Common Crawl data set."

During their recorded discussions, Meta executives talked about how they had hired contractors in Africa to aggregate summaries of fiction and nonfiction. The summaries included copyrighted content "because we have no way of not collecting that," a manager said in one meeting.

Meta's executives said OpenAI seemed to have used copyrighted material without permission. It would take Meta too long to negotiate licenses with publishers, artists, musicians and the news industry, they said, according to the recordings.

"The only thing that's holding us back from being as good as ChatGPT is literally just data volume," Nick Grudin, a vice president of global partnership and content, said in one meeting.

OpenAI appeared to be taking copyrighted material and Meta could follow this "market precedent," he added.

Meta's executives agreed to lean on a 2015 court decision involving the Authors Guild versus Google, according to the recordings. In that case, Google was permitted to scan, digitize and catalog books in an online database after arguing that it had reproduced only snippets of the works online and had transformed the originals, which made it fair use.

Using data to train A.I. systems, Meta's lawyers said in their meetings, should similarly be fair use.

At least two employees raised concerns about using intellectual property and not paying authors and other artists fairly or at all, according to the recordings. One employee recounted a separate discussion about copyrighted data with senior executives including Chris Cox, Meta's chief product officer, and said no one in that meeting considered the ethics of using people's creative works.

'Synthetic' Data

OpenAI's Mr. Altman had a plan to deal with the looming data shortage.

Companies like his, he said at the May conference, would eventually train their A.I. on text generated by A.I. — otherwise known as synthetic data.

Since an A.I. model can produce humanlike text, Mr. Altman and others have argued, the systems can create additional data to develop better versions of themselves. This would help developers build increasingly powerful technology

and reduce their dependence on copyrighted data.

“As long as you can get over the synthetic data event horizon, where the model is smart enough to make good synthetic data, everything will be fine,” Mr. Altman said.

A.I. researchers have explored synthetic data for years. But building an A.I system that can train itself is easier said than done. A.I. models that learn from their own outputs can get caught in a loop where they reinforce their own quirks, mistakes and limitations.

“The data these systems need is like a path through the jungle,” said Jeff Clune, a former OpenAI researcher who now teaches computer science at the University of British Columbia. “If they only train on synthetic data, they can get lost in the jungle.”

To combat this, OpenAI and others are investigating how two different A.I. models might work together to generate synthetic data that is more useful and reliable. One system produces the data, while a second judges the information to separate the good from the bad. Researchers are divided on whether this method will work.

A.I. executives are barreling ahead nonetheless.

“It should be all right,” Mr. Altman said at the conference.

Read by Cade Metz

Audio produced by Patricia Sulbarán.

A correction was made on April 6, 2024*: An earlier version of this article misstated the publisher of J.K. Rowling’s books. Her works have been published by Scholastic, Little, Brown and others. They were not published by Simon & Schuster.*

Cade Metz writes about artificial intelligence, driverless cars, robotics, virtual reality and other emerging areas of technology. More about Cade Metz

Cecilia Kang reports on technology and regulatory policy and is based in Washington D.C. She has written about technology for over two decades. More about Cecilia Kang

Sheera Frenkel is a reporter based in the San Francisco Bay Area, covering the ways technology impacts everyday lives with a focus on social media companies, including Facebook, Instagram, Twitter, TikTok, YouTube, Telegram and WhatsApp. More about Sheera Frenkel

Stuart A. Thompson writes about how false and misleading information spreads online and how it affects people around the world. He focuses on misinformation, disinformation and other misleading content. More about Stuart A. Thompson

Nico Grant is a technology reporter covering Google from San Francisco. Previously, he spent five years at Bloomberg News, where he focused on Google and cloud computing. More about Nico Grant

---

A version of this article appears in print on , Section A, Page 1 of the New York edition with the headline: How Big Tech Cut Corners to Harvest Data for Their A.I. Models